RESEARCH INITIATIVE FOR SECURE AI-POWERED DECISION-MAKING

PROBLEM

Competitive pressures force more and more outsourcing of decision-making to AI. However, it is unclear whether the models inherit or resist human reactions to **"judgment-impairing conditions"** such as time-pressure, information ambiguity, high-stakes situations, and more.

We need a better understanding of how models maintain a standard for reasoning and how humans can engage with their output under such conditions. Otherwise, humanity risks gradually losing control over incresingly AI-powered decision-making.

RESEARCH AIM

Uncover harmful weaknesses in decision-making between people and AI in situations where judgement is impaired.

RESEARCH QUESTIONS

Technical: How can we measure the robustness of model performance under judgment-impairing conditions?

Cognitive: How do such conditions impact human decision-makers' overreliance and how transparent is this risk to them?

Remedial: How can the identified vulnerabilities be used to shape model evaluations, early-warning alerts, and safeguards?

IMPACT

A foundation for monitoring the integrity of human choice and enable the development of safeguards against disempowerment in areas facing the most immediate risk.

STAGES ///

1

Model reasoning evaluation

Define how model judgment change under judgment-impairing conditions. Develop appropriate evaluation protocols.

2

Human-AI dynamics mapping

Identify factors enhancing overreliance on models for assistance in contexts where human ability to oversee decisions is impaired.

3

Exploring safeguards

Describe failure modes of human-AI collaboration on decision-making. Create monitoring tools for the integrity of choice and derive adequate safeguards.

4

Real-life implementations

Adjust monitoring and safeguards for the needs of real-world actors who currently implement AI in decision-making processes.

Does this sound interesting? Reach out. Seeking both researchers and operations talent.

Nowe Moore | ndatkova@sas.upenn.edu | Nowe Moore | nowemoore.com